DOCUMENT RESUME

ED 450 152 TM 032 342

AUTHOR French, Christine L.
TITLE A Review of Classical Methods of Item Analysis.
PUB DATE 2001-02-00
NOTE 22p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Item Analysis; Item Response Theory; *Selection; *Test Items

ABSTRACT

Item analysis is a very important consideration in the test development process. It is a statistical procedure to analyze test items that combines methods used to evaluate the important characteristics of test items, such as difficulty, discrimination, and distractibility of the items in a test. This paper reviews some of the classical methods for item analysis. The paper also provides an explanation of common item analysis procedures, which aid in the selection of appropriate test items. In addition, the major limitation of some of the classical methods of item analysis and some benefits of other methods, such as Item Response Theory, are discussed. (Contains 1 table and 12 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running Head: METHODS OF ITEM ANALYSIS

ED 450 152

A Review of Classical Methods of Item Analysis

Christine L. French

Texas A&M University 77843-4225

C. L. French

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM032342

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 1-3, 2001.

BEST COPY AVAILABLE

Abstract

Item analysis is a very important consideration in the test development process. The present paper reviews some of the classical methods for item analysis. The paper also provides an explanation of common item analysis procedures, which aid in the selection of appropriate test items. In addition, the major limitation of the classical methods of item analysis and some benefits of other methods, such as Item Response Theory, will be discussed.

A Review of Classical Methods of Item Analysis

There are many different facets of test development. One must go through a series of steps in order to create a test that best suits the test developer's purposes. These steps include test conceptualization, test construction, test tryout, analysis, and revision (Cohen, Swerdlick, & Phillips, 1996). By the time a test developer has designed a plan and purpose for a specific test, written items, and tried out the initial test in a sample, there is still much work left to be done. Much effort and time must be spent with an analysis of the items. Although it is only one part of the test development process, item analysis is a key step that determines the psychometric properties of the test in question. As such, educators, academicians, and others must not take the item analysis and selection process lightly.

There is more than one way to perform an analysis of the items of a test. However, for the purposes of this paper, only the "classical" methods of item analysis and selection will be discussed. The primary objective of the paper is to review item analysis and item selection. This discussion will include a summary of item analysis, a review of the purpose and goals of the item analysis process, the general methods of classic item analysis, and the major limitation of classical item analysis and selection methods. In addition, the paper will briefly

discuss other non-classical or "modern" methods of item analysis.

Summary of Item Analysis

Item analysis, in short, is a statistical procedure conducted to analyze test items (Cohen, Swerdlick, & Phillips, 1996). It is a combination of methods used to evaluate the important characteristics of test items, namely the difficulty, discrimination, and distractibility of the items in a test (Hills, 1981; Thompson & Livitov, 1985). In addition, item analysis is used to examine the validity and reliability of scores on test items (Cohen, Swerdlick, & Phillips, 1996; Gall, Borg, & Gall, 1996).

As suggested, part of the item analysis process is quantitative in nature. However, there also is a qualitative aspect to item analysis. These methods include talking with the examinees in the tryout sample about their thoughts, problems, and comments concerning the test. A test developer also may want to evaluate the responses given by the examinees to gain further qualitative information (Johnson, 1977). These methods can reveal important information that would otherwise not be evident by a quantitative item analysis alone. Problems with cultural sensitivity, face validity, test fairness, and length cannot be gathered from a pure statistical analysis of the items (Cohen, Swerdlick, & Phillips, 1996). Therefore, test developers should

be sure to have a good balance of quantitative and qualitative methods to analyze the items in the test. For a further description of qualitative analysis methods, see Cohen, Swerdlick, & Phillips (1996).

Item analysis is not confined to one type of test. Rather, this statistical method to aid in choosing the most appropriate items for a specific test can be applied to achievement and personality tests, as well as to behavioral and self-report inventories. However, item analysis is limited to use with objective test items (Johnson, 1977). Specifically, there must be a defined set of possible answers for a test item in order to gather data on the item's psychometrics. If item analysis were applied to an essay question, there would be limitless answers possible and an analysis of this type of item would result only in the psychotic break of the test developer.

Purpose and Goals

As mentioned before, the test development process includes five different steps, including test conceptualization, test construction, test tryout, analysis, and revision (Cohen, Swerdlick, & Phillips, 1996). As can be seen, item analysis follows an initial tryout of the test. The order of the steps included in the test development process imply the purpose of item analysis, which is to evaluate and revise a test that has been given to a sample of examinees.

As a recently developed test is evaluated and revised, there is an underlying goal behind the statistical analyses that are being performed. The goal of item analysis methods is to maximize the psychometric quality of scores from the test (Anastasi & Urbina, 1997; Ebel & Frisbie, 1986). Evaluating the difficulty and discrimination of the test items are two ways of controlling the psychometric properties of the items on a test (Ebel & Frisbie, 1986).

General Methods

There are many different item analysis methods, but only a few will be covered here. The methods to be discussed include the following item indices: difficulty, discrimination, and distractibility.

Item Difficulty Index

The item difficulty index, or p, is a proportion of the number of examinees who get an item correct to the total number of examinees (Anastasi & Urbina, 1997; Cohen, Swerdlick, & Phillips, 1996). Therefore, the easier an item, the larger the proportion will be. For example, a math problem that is answered correctly by 22% (p = .22) of a class would be considered harder than a math problem that is answered correctly by 82% of a class (p = .82). In general, the item that has a difficulty index of .22 would be considered rather difficult and some might even assert that the concept covered by the item needs more

discussion (Hills, 1981). However, this same item (p = .22) might not need more discussion at all. It may need to be revised or eliminated for any number of possible reasons (Hills, 1981).

The item difficulty index ranges from zero to one. However, if p approaches either end of the spectrum (0 or 1.0) for a particular test item, less information is revealed about a group of examinees because the test item does not differentiate between the high scorers and the low scorers (Anastasi & Urbina, 1997; Johnson, 1977)). Likewise, as p approaches the midpoint (.50), more information is revealed because the high and low scoring groups are differentiated (Anastasi & Urbina, 1997; Johnson, 1977).

Every item on a test has its own item difficulty index, or p. However, test developers, teachers in particular, may want to evaluate the overall item difficulty. To get an index of the average test item difficulty for an entire test, one must find the mean of the item difficulty indices for all the items on the test. The average test item difficulty can be represented by the following formula:

$$p_{total} = (\Sigma\, p_i) \;/\; n$$

For example, Mrs. Jones has developed a history test that has five questions (n = 5), with item difficulty indices of p_1 = .25, p_2 = .37, p_3 = .48, p_4 = .64, p_5 = .89. Using the aforementioned formula, p_{total} = .53 for this test. According to Cohen,

Swerdlick, and Phillips (1996), the "optimal average item difficulty is approximately .5" (p. 234), at least on a "supply-format" item (e.g., fill-in-the-blank). Therefore, Mrs. Jones can feel confident that the average difficulty of the items on this test are at an appropriate level for her second period history class, trusting that her initial sample of students was similar to the students currently in her class.

However, there is a caveat to the index of average test item difficulty. As with all tests, there will be some guessing on the part of the examinees, if a "selection" format is used. With this in mind, the desired proportion of correct responses should be set higher than .5, because this percentage of correct answers can be achieved on the basis of random guessing alone (Anastasi & Urbina, 1997; Cohen, Swerdlick, & Phillips, 1996). Accordingly, the best index of average item difficulty should be set at the midpoint between 1.0 and the chance responding percentage (Anastasi & Urbina, 1997; Cohen, Swerdlick, & Phillips, 1996). Returning to Mrs. Jones' history test, assume each of the five questions has four choices. In this example, the best item difficulty should be approximately .63 because the probability of randomly guessing correctly on any item is .25 (Cohen, Swerdlick, & Phillips, 1996).

Item difficulty does not always measure difficulty in its true sense, but rather endorsement of an item (Cohen, Swerdlick,

& Phillips, 1996). For tests of personality or behavior, as the item difficulty index approximates 1.0, the greater number of people who endorse that particular item.

Overall, test items should be tailored to the specific objectives of instruction. The item difficulty is always left up to the test developer, who then considers what purpose the test is going to serve and the ability of the group being tested (Hambleton, 1993). It is possible that Mrs. Jones simply wants to evaluate the mastery level of her history students. In this case, she would want to use a test with an average item difficulty that is slightly higher (more students get more items correct) than her calculated optimal item difficulty index of $\underline{p}$ = .63 (Johnson, 1977).

Item Discrimination Index

The item discrimination index, or $\underline{d}$, refers to how well an item is able to differentiate high and low scoring test takers (Anastasi & Urbina, 1997; Cohen, Swerdlick, & Phillips, 1996). The item discrimination index ranges from -1.0 to +1.0 and is an indication of the high and low scoring students who chose the correct answer on a particular test item (Hills, 1981). A high index of item discrimination ($\underline{d}$ > .40) will always preferred over a lower index of discrimination (Ebel & Frisbie, 1986). In order to find the item discrimination index for a particular item, consider the following formula:

$$\underline{d} = (U - L) / n$$

The item discrimination index is equal to the number of students in the upper scoring group, U, minus the number of students in the lower scoring group, L, who get the correct answer on a certain question. The difference is then divided by the total number of students in each group (Cohen, Swerdlick, & Phillips, 1996).

For example, Mrs. Jones wants to give her second period history class the five-question test she recently developed. Being a very large high school, there are one hundred students in her class. Assuming the distribution of scores is normal, Mrs. Jones isolates the top and bottom 27% of the students who took the test (Cohen, Swerdlick, & Phillips, 1996).

INSERT TABLE 1 ABOUT HERE.

As shown on Table 1, an example of perfect item discrimination is evidenced by item four on Mrs. Jones' test. All of the 27 students in the upper scoring group got item four correct and none of the 27 students in the lower group got item four correct, therefore $\underline{d} = 1.0$. An example of an item that does not discriminate between groups is evidenced by item three on the history test. The same number of students in the upper group as the lower group got the correct answer, therefore $\underline{d} = 0$. This

item would need to undergo tremendous revision in order to discriminate between the upper and lower scoring groups. As suggested, as $\underline{d}$ approaches 1.0, the better the item discrimination; as $\underline{d}$ approaches 0, the degree to which an item is able to discriminate lessens (Cohen, Swerdlick, & Phillips, 1996).

Consider item five on Mrs. Jones' test. The index of discrimination is a negative number, which means that more people in the lower group than the upper group got the item correct. As some might say (Cohen, Swerdlick, & Phillips, 1996), this situation is a "test developer's nightmare" (p. 238) and the item needs to be revised or thrown out completely. This situation might have been occurred because item five was worded so vaguely that the upper scoring students were making inferences and interpretations that precluded them choosing the correct answer (Johnson, 1977).

It might seem that the index of discrimination is a rather subjective number to which a test developer places their own value on the number derived. However, there is a general rule about the preference level for an item discrimination index. Anastasi and Urbina (1997) suggested a level above or as close to 50% as possible. Others have laid out a guideline of all the possible discrimination index values and their evaluation. Ebel and Frisbie (1986) suggested that item discrimination indices

greater than .40 are very good items, those between .30 and .39 are good but there is some room for revision, those between .20 and .29 are borderline and are in need of improvement, and those below .19 should be eliminated or undergo much improvement (p. 234).

There are two particular useful characteristics of the item discrimination index compared to other correlations, such as point-biserial, biserial, Flanagan's, and Davis' correlations (Ebel & Frisbie, 1986). As Ebel and Frisbie (1986) state, d is "simpler to compute and explain to others" and "it has the very useful property…of being biased in favor of items of middle difficulty" (pp. 229-230).

The index of discrimination is essentially a test developer's tool to be able to see how well the items of a test are able to discriminate between high and low scoring examinees. However, the item discrimination index is highly subject to sampling error (Ebel & Frisbie, 1986). For example, if a test is given to two groups of examinees, one to a very homogenous group and another that is more extreme in their characteristics which the test wants to discriminate, the index of discrimination for some of the items will differ depending on the sample. This is one of the limitations of such classical item analysis methods and will be discussed later in the paper.

Item Distractor Index

The item distractor index is relatively less used when compared to item difficulty and item discrimination indices. On a general level, it represents a combination of these two more popular indices (p and d), in addition to an evaluation of the pattern of responses to any of the extraneous, decoy responses for a particular item (Hills, 1981). Similarly, if previous analysis has evidenced a good level of average test difficulty (approximately .50) and high discrimination values (greater than .40), there is little need to perform such an analysis of the distractors in an item (Hills, 1981).

There are three essential rules on which the item distractor index is based, as suggested by Hills (1981). First, every option is chosen by at least one examinee. Second, more students in the upper scoring group will choose the right answer than those in the lower scoring group. Third, the reciprocal of this second tenet is that more students in the lower scoring group will select the wrong choices, also known as decoys or distractors, more often than those in the upper scoring group. As can be hypothesized, the item distractor index might possibly contribute much information to the qualitative analysis of the development of a test and the selection of items.

Limitation of Classical Methods

As previously mentioned, there is a major limitation with regard to the classical methods of item analysis. This limitation is the result of inseparable examinee and test characteristics.

Not only do the criteria for the best items for a test differ depending on the objectives of the test developer (Cohen, Swerdlick, & Phillips, 1996), the criteria for items also differ depending on the samples used for the tryout version of a test. In other words, the previously described methods that are touted to ensure the test developer challenging and reliable test scores, are the very methods that are partially based on sampling error. As Ebel and Frisbie (1986) noted, the index of item discrimination is highly subject to sampling error. This premise also would extend to the indices of item difficulty and item distractibility.

In classical item analysis and development methods, the examinee and test characteristics are inseparable (Hambleton, Swaminathan, & Rogers, 1991). Consider Mrs. Jones' history test once again. After developing the five-item test, she tested it on her second period class just a few days prior to summer vacation. She then calculated the different item indices and was pleased with the level of difficulty and discrimination of each of the items. Once the students returned from summer break, she

gave the test to her new second period class on the second day of school in the fall. This new group of students is different than the tryout sample of students in terms of age and the bulk of knowledge about the subject being tested. Most likely, the items will seem more difficult to the new students. In addition, the items will probably not have the same level of discrimination as they did when the test was first administered to the other students before they left for summer break.

Overall, the main limitation of classical item analysis methods is that the methods are group-dependent and change depending on the group being tested. Success of classical item analysis methods "depends directly on how closely the sample used to determine the item parameters employed in the item-selection process matches the population for which the test is intended" (Hambleton, 1993, p. 184). Similarly, Hambleton, Swaminathan, and Rogers (1991) asserted that "group-dependent item indices are of limited use when constructing tests for examinee populations that are dissimilar to the population of examinees with which the item indices were obtained" (p. 3).

Other Methods of Analysis

There are some other ways in which a test developer can assess the items in a newly created test. Most of these methods are done by computers and have several inherent advantages. Two

of the main advantages include the accuracy and the elaboration of the results (Hills, 1981).

One specific method of item analysis is called Item Response Theory (IRT), of which there are three primary models. IRT is also known as latent trait theory Gall, Borg, & Gall, 1996). According to Hulin, Drasgow, and Parsons (1983), IRT takes into consideration the fact that an individual's response on a test is somehow related to an underlying characteristic. One of the main benefits of IRT is the unbiased nature of the results that are not dependent on a particular sample (Anastasi & Urbina, 1997).

For the most part, computer-adapted tests (such as recent versions of the GRE, GMAT, etc.) are based on item response theory. The scores yielded by such analyses are more reliable and sample-free and facilitate the construction of parallel test forms that are truly equal (Gall, Borg, & Gall, 1996). Furthermore, IRT selects items in a more direct fashion than is possible with classical item analysis methods, with a focus on using the test information curve (TIC) instead of the item difficulty index or correlation coefficients (Hulin, Drasgow, & Parsons, 1983). However, IRT is not magical, as explained in some detail by Fan (1998) and Lawson (1991).

Conclusion

Classical item analysis and selection methods remain popular means of assessing and evaluating newly developed tests, particularly within the classroom setting. It is important for a test developer or researcher to be familiar with the different methods in order to facilitate the development of a new test. It is also important for the users of these tests to understand these methods and their limitations, which will help them to evaluate the quality of the tests.

Items are selected based on several different facets, all of which contribute to the total amount of information that can be supplied to the test developer by the test as a whole (Hulin, Drasgow, & Parsons, 1983). These facets of item analysis include the difficulty and discrimination of an item, as well as the distractibility of the possible choices (Hills, 1981; Hulin, Drasgow, & Parsons, 1983). Essentially, the criteria for the best items depend solely on the objectives and values of the person developing the test itself (Cohen, Swerdlick, & Phillips, 1996).

This paper has presented the general methods of classical item analysis and selection. Test developers must take into consideration the objectives of the test and their own values with regard to item indices. Considering all of this, the test

developer can then determine the best route of item exploration, be it classical or computer-based IRT methods.

References

Anastasi, A., & Urbina, S. (1997). Item analysis. In Psychological testing (7th ed., pp. 172-202). Upper Saddle River, NJ: Prentice-Hall.

Cohen, R.J., Swerdlick, M.E., & Phillips, S.M. (1996). Test development. In Psychological testing and assessment: An introduction to tests and measurement (3rd ed., pp. 218-254). Mountain View, CA: Mayfield.

Ebel, R.L., & Frisbie, D.A. (1986). Using test and item analysis to evaluate and improve test quality. In Essentials of educational measurement (4th ed., pp. 223-242). Englewood Cliffs, NJ: Prentice-Hall.

Fan, X. (1998). Item Response Theory and classical test theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement, 58, 357-381.

Gall, M.D., Borg, W.R., & Gall, J.P. (1996). Educational research: An introduction (6th ed.). White Plains, NY: Longman.

Hambleton, R.K. (1993). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 147-200). Phoenix, AZ: Oryx.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Hills, J.R. (1981). Measurement and evaluation in the classroom (2nd ed.). Columbus, OH: Charles E. Merrill.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Methods. In Item response theory: Application to psychological measurement (pp. 75-109). Homewood, IL: Dow Jones-Irwin.

Johnson, M.C. (1977). A review of research methods in education. Chicago: Rand McNally College.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 159-168). Greenwich, CT: JAI Press.

Thompson, B., & Livitov, J.E. (1995). Using microcomputers to score and evaluate test items. Collegiate Microcomputer, 3, 163-168.

Table 1

Index of Item Discrimination for Mrs. Jones' History Test

Item	Upper	Lower	Upper - Lower	n	d
1	20	5	15	27	.56
2	17	15	2	27	.07
3	12	12	0	27	0
4	27	0	27	27	1.0
5	5	12	-13	27	-.48